UCRL-CR-142043
B510060

Final Report: Spatio-Temporal Data Mining of Scientific Trajectory Data

S. Gaffney, P. Smyth

January 10, 2001

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

DISCLAIMER

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

University of California, Irvine
414E Computer Science
Information and Computer Science
CA, 92697-3425

Final Report for Spatio-Temporal Data Mining of Scientific Trajectory Data

Submitted by:
PI: Padhraic Smyth
Associate Professor

FINAL REPORT
For the period ending September 30th 2000

Prepared for:
University of California
Lawrence Livermore National Laboratory
Attn: Jill Dunaway
P.O. Box 808, L-561
Livermore, CA 94551

Under
B510060

Date prepared
January 10th 2001

FINAL REPORT: Spatio-Temporal Data Mining of Scientific Trajectory Data

Scott Gaffney & Padhraic Smyth
Department of Information and Computer Science
University of California, Irvine
Irvine CA 92697-3425
{sgaffney, smyth}@ics.uci.edu

January 10, 2001

1 Introduction and Background

With the increasing availability of massive observational and experimental data sets (across a wide variety of scientific disciplines) there is an increasing need to provide scientists with efficient computational tools to explore such data in a systematic manner. For example, techniques such as classification and clustering are now being widely used in astronomy to categorize and organize stellar objects into groups and catalogs, which in turn provide the impetus for scientific hypothesis formation and discovery (e.g., see Fayyad, Djorgovski and Weir (1996); or Cheeseman and Stutz (1996) or Fayyad and Smyth (1999) in a more general context).

Data-driven exploration of massive *spatio-temporal* data sets is an area where there is particular need of data mining techniques. Scientists are overwhelmed by the vast quantities of data which simulations, experiments, and observational instruments can produce. Analysis of spatio-temporal data is inherently challenging, yet most current research in data mining is focused on algorithms based on more traditional feature-vector data representations.

Scientists are often not particularly interested in raw grid-level data, but rather in the phenomena and processes which are "driving" the data. In particular, they are often interested in the temporal and spatial evolution of specific "spatially local" structures of interest, e.g., birth-death processes for vortices and interfaces in fluid-flow simulations and experiments, trajectories of extra-tropical cyclones from sea-level pressure data over the Atlantic and Pacific oceans, and sunspot shape and size evolution over time from daily chromospheric images of the Sun. The ability to automatically detect, cluster, and catalog such objects in principle provides an important "data reduction front-end" which can convert 4-d data sets (3 spatial and 1 temporal dimension) on a massive grid to a much more abstract representation of local structures and their evolution. In turn, these higher-level representations

provide a general framework and basis for further scientific hypothesis generation and investigation, e.g., investigating correlations between local phenomena (such as storm paths) and global trends (such as temperature changes).

In this work we focused on detecting and clustering *trajectories* of individual objects in massive spatio-temporal data sets. There are two primary technical problems involved. First, the local structures of interest must be detected, characterized, and extracted from the mass of overall data. Second, the evolution (in space and/or time) of these structures needs to be modeled and characterized in a systematic manner if the overall goal of producing a reduced and interpretable description of the data is to be met.

Existing data-mining and statistical tools for clustering and classification are largely based on the so-called *feature-vector representation* of an object. For example, for a stellar object one can measure characteristics such as brightness, shape, size, and so forth. Given p such measurements we can think of the p-dimensional measurement vector as existing in a p-dimensional Euclidean space. Notions such as distance, similarity, decision boundaries, prototypes, clusters, and so forth, all have a natural geometric notion in such a representation. Indeed, it is safe to say that the vast majority of classification and clustering techniques which currently exist in data analysis are cast in this *multivariate* framework.

Dynamic trajectories of objects are difficult to handle with traditional vector-based clustering methods (e.g., representing a sequence of (x_i, y_i), $1 \leq i \leq n$ position measurements as a vector of length $2n$). The main difficulties are:

- different objects have trajectories of different lengths and may evolve at different time-scales, making a vector description inappropriate.
- object trajectories are inherently smooth as a function of time, information which is lost by vectorization.
- objects have additional features of interest such as size, (e.g., sunspot size), shape, and velocity as well as their 2d or 3d position as a function of time. One would like to be able to systematically model the interdependence of the object's spatial evolution and its features.

Our proposed work involved development of general techniques for (a) extracting trajectories of moving objects from large data spatio-temporal data archives, and (b) developing techniques for clustering such spatio-temporal trajectories (e.g., as in Gaffney and Smyth, 1999). Most of our actual work during the award period was devoted to the *detection* and *tracking* problems, specifically the development of algorithms and software to extract 2d-dimensional cyclone paths from sea-level pressure spatial data records.

2 Overview of Work Performed

In this report we summarize what has been done this year for this project. Below we describe the main points and give some further details in the following sections.

- Section 3: Research into dynamic modelling
  - Generated trajectory-like sequences from AR, MA, and ARMA models
  - Investigated how much data is needed to learn the parameters of these models
  - Investigated the feasibility of learning a mixture of these models
- Section 4: Experimentation with application to cyclone clustering
  - Obtained an appropriate meteorlogical data set
  - Finished extensive preprocessing of the data for our purposes
  - Chose a bicubic interpolation and gradient descent based method to enable offgrid cyclone tracking.
  - Performed some basic cyclone tracking experiments with the newly developed software.
  - Developed a simple GUI to enable the visualization of the tracking results.
- Section 5: Software development
  - Finished the porting of our previous cyclone tracking software
  - Finished the modifications to enable offgrid tracking
- Section 6: Future work
  - Further investigate mixtures of dynamic models, including both AR models as well as Kalman filter models
  - Investigate ways in which the tracking of cyclones (or other phenomenon) can be integrated into their clustering
  - Obtain another data set in a different application area from cyclone clustering

3 Research Into Dynamic Modelling

We proposed a general framework for clustering trajectories using probabilistic models of dynamic systems which allows one to overcome limitations of feature-vector based methods. As such we began looking at various types of dynamic models, for example, autoregressive (AR), moving average (MA), and the more general ARMA model.

One of our tasks was to figure out how to simulate direction-focused trajectories from these models in terms of parameter settings. Figure 1 shows some simple trajectories generated from an AR model. These sequences have a loose direction of travel instead of randomly moving about. Its this type of behavior that we are trying to model from scientific data sets.

We also looked at learning the parameters of these models given some set of generated data from a known model. We investigated exactly how much data one

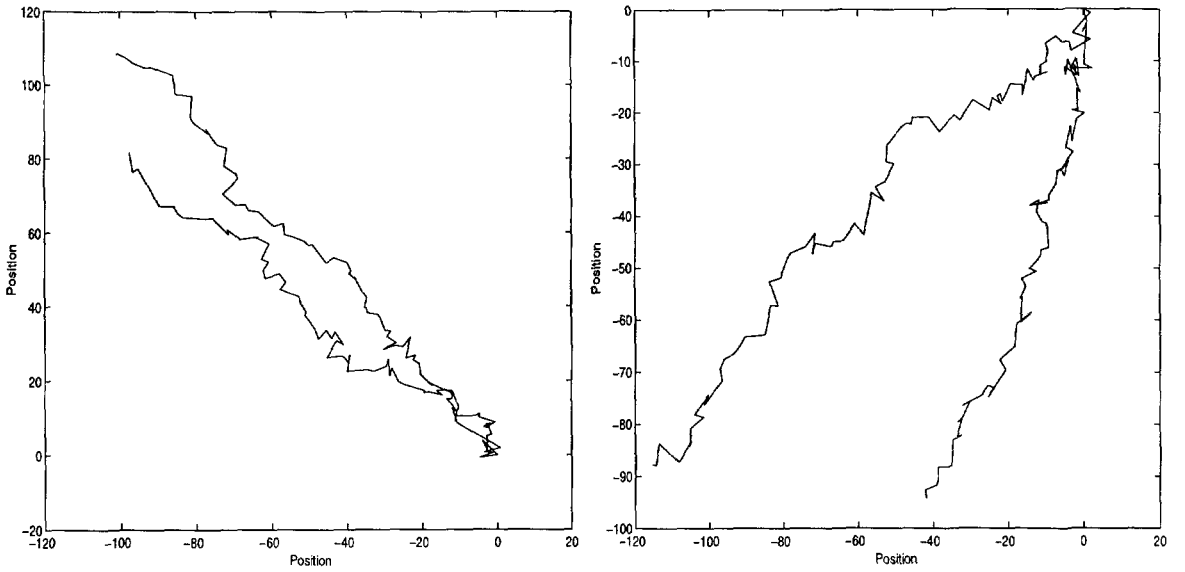

Figure 1: Two different pairs of sequences generated from an AR(2) process. Each sequence begins at (0,0).

needs to accurately learn a good model. Of importance is how difficult it is to learn a mixture of AR models, since mixture models are at the core of our trajectory clustering technique. We are currently looking deeper into this issue.

4 Experimentation with Application to Cyclone Clustering

One of the applications of trajectory clustering can be found in the clustering of cyclone tracks from meteorological data (e.g., Hodges, 1994; Blender et al., 1997). We are currently working with the CCM3 AMIP II simulated data set for the 1979/1980 winter that gives mean sea-level pressure (MSLP) measurements on a $2.5° \times 2.5°$ grid over the earth. In order for the data to be usable for our purposes we had to perform some extensive preprocessing to filter out long term effects in the measured field. A snapshot of the resulting data can be seen in Figure 2.

In a previous project update we noted that an important aspect of our proposed approach of employing dynamic models for cyclone tracking (in this case) is that we somehow need to be able to track cyclones in continuous state-space. This means that we cannot be confined to the aforementioned grid. As such we must employ some kind of interpolation scheme so that we can track cyclones off of the grid.

Here we focus on using a bicubic interpolation inside of an iterative scheme to find our minima using a simple gradient descent. First we scan all of the images (or the MSLP data slices over time) and find all the local minima using a simple sliding neighborhood method. That is, we declare a "pixel" to be at a local minimum if its value is less than all eight of its neighbors. Then we use a simple gradient descent with bicubic interpolation to descend to the point "inside" of the pixel that is at an

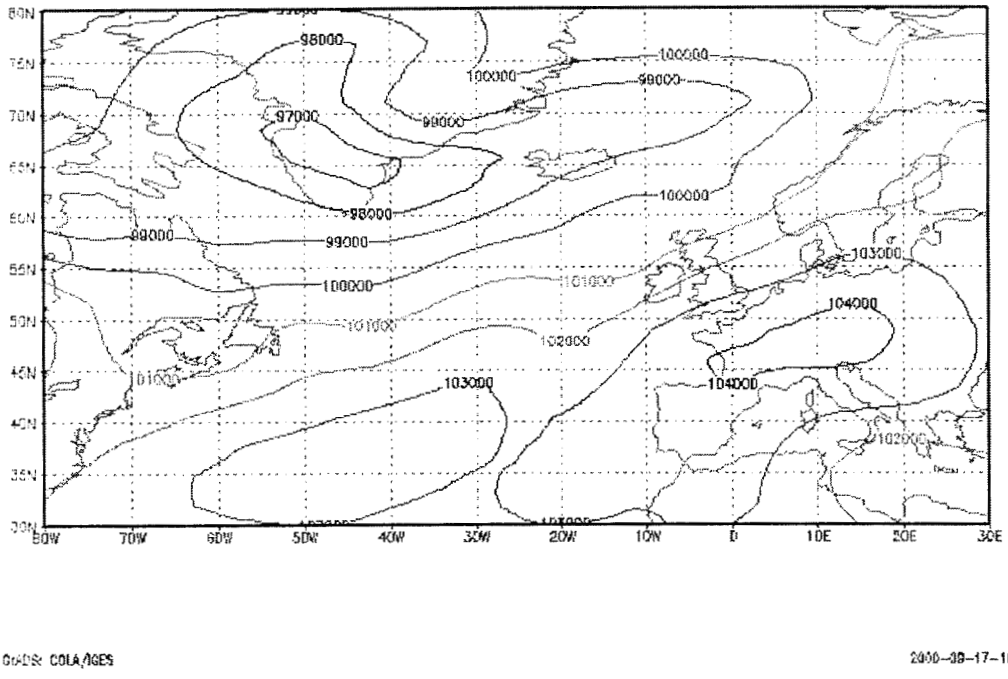

Figure 2: Contour plot showing MSLP in the North Atlantic at a particular date.

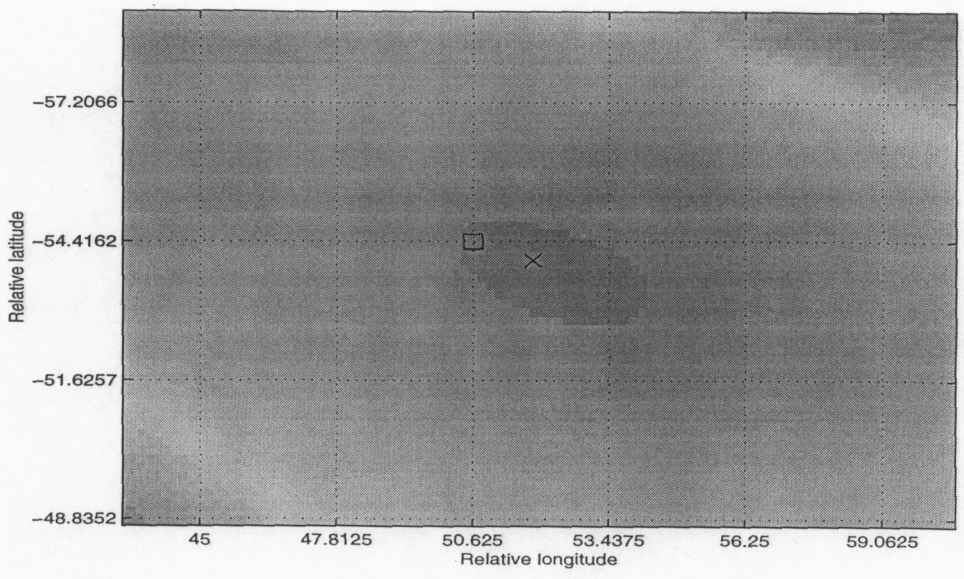

Figure 3: An example offgrid minimum found using gradient descent. We see an image of the interpolated MSLP data at one instant in time. The grid lines represent the location of the acutal data grid in our data set. The square shows the grid-located minimum found using sliding neighborhoods. The 'x' shows the approximate minimum found using gradient descent with bicubic interpolation.

approximate minimum. This point then gives us our approximate offgrid center of a candidate cyclone. Figure 3 shows an example of an offgrid minimum that was found using this method. The image shown has been interpolated so that we can see "inside" of each pixel.

Using the above technique, we processed the data to force all of the grid-based minima to lie in continuous space. We then fed this data into our new tracking software and observed the results. Figure 4 shows an example of a single trajectory that was generated from the above steps. The image shown displays the MSLP data at the instant in time when the cyclone is at the far right of its trajectory. A simple GUI was developed so that we could visualize the results and determine not only the performance of the tracking, but also to get a grasp on the difficulty of the problem at hand.

5 Software Development

We have finished the process of porting our previously developed MATLAB software to our current C++ PC-based platform and, in addition, we have completed the necessary modifications to allow tracking to be carried out using offgrid coordinates. At this time, much of the basic software development has been finished. Much of the future development will be focused on dynamic modelling implementation.

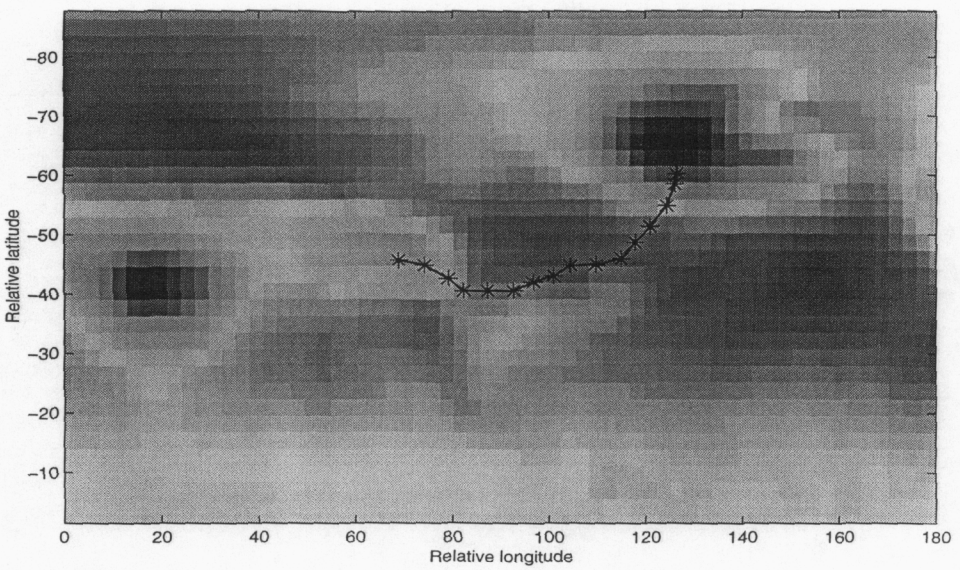

Figure 4: A single generated cyclone trajectory. The line in the figure connects the '*' symbols along the trajectory that represent the minima that were tracked as part of this cyclone. The background image represents the MSLP data at the time the cyclone is centered at the far right '*'.

6 Future Work

As planned, now that we have finished most of the basic software development, we can focus more on the modelling aspect of the project. Our goal is to target the investigation of Kalman filter models (e.g. North and Blake, 1998) and their application to this project. Since Kalman filters operate in continuous space, our new tracking software will complement this task nicely.

From here we would like to investigate ways in which the tracking can be integrated into the clustering framework (e.g., Blender et al. (1997)). That is, if we know which cluster an individual belongs to with some probability, then we should be able to more accurately track his future movements. In other words, we believe that instead of two different problems—tracking and clustering—what we have here is one compound problem that can be solved in an integrated manner.

References

Blender, R., Fraedrich, K., and Lunkeit, F. (1997) Identification of cyclone-track regimes in the North Atlantic. *Quart J. Royal Meteor. Soc.*, 123, 727–741.

Cheeseman, P. and Stutz. J. (1996) Bayesian classification (AutoClass): theory and results. In *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.), Cambridge, MA: AAAI/MIT Press, pp. 153–180.

Fayyad U.M., Djorgovski S.G., and Weir N. (1996) Automating the analysis and cataloging of sky surveys, In *Advances in Knowledge Discovery and Data Mining*, ed. U.M.Fayyad, G.Piatetsky-Shapiro, P.Smyth, and R.Uthurusamy, Menlo Park, California: AAAI Press. 471-493.

Fayyad, U. M. and Smyth, P. (1999) Cataloging and mining massive databases for science data analysis, *Journal of Graphics and Computational Statistics*, 8(3), 589–610.

Gaffney, S. and Smyth P. (1999) Trajectory clustering with mixtures of regression models. In *Proceedings of the 1999 ACM Conference on Knowledge Discovery and Data Mining.*

Hodges, K. I. (1994) A general method for tracking analysis and its application to meteorological data. *Mon. Wea. Rev.*, 122, 2573-2586.

North, B. and Blake, A. (1998) Learning dynamical models by expectation maximization. *Proceedings of the 6th International Conference on Computer Vision.*